Team and Collective Performance Measurement

Ebb Smith
DSTL, Policy and Capability Studies
Bedford Technology Park
Thurleigh, Bedfordshire
UNITED KINGDOM

Email: mesmith@dstl.gov.uk

Jonathan Borgvall, Patrick Lif
Swedish Defence Research Agency
Department of Man-System Interaction
SE-581 11 Linkoping
SWEDEN

Email: {jonathan.borgvall, patrik.lif} @foi.se

1.0 INTRODUCTION

The measurement of operator performance has for some time formed the basis of research for those engaged in the field of human system interaction and the use of virtual reality (VR). Performance measurement is particularly relevant when the desire is develop methods and metrics to assess the utility of VR for training purposes and to predict how well that training will then transfer to the real world. Performance measurement becomes even more critical when the VR application is used in a military context, e.g., in preparation for conflict.

This chapter provides descriptions of some of the methods and measures used for measuring task and mission performance in virtual environments. As one of the challenges inherent in assessment of VR is the measurement of team and collective performance, this is the primary focus of the chapter.

2.0 TEAM PERFORMANCE

A team performance measurement system must be able to distinguish between individual and team level performance deficiencies, i.e., both *taskwork* and *teamwork* behaviours [1]. Taskwork behaviours are those performed by individual team members to execute their specific functions, e.g., weapons systems switchology. Teamwork behaviours are those which are related to team member interactions and the co-ordination of team members to achieve a common goal, e.g., communication, compensatory behaviours, information flow and feedback. For example, a team may make an incorrect decision because information was not circulated effectively among the team members (a team level problem). However, the same incorrect decision could be made because an individual made a technical error, which is an individual level problem [2].

A measurement system should assess both team *outcomes* and team *processes*. Outcomes are the end result of team performance (e.g., mission effectiveness – number of targets hit) and processes are the specific behaviours and performance strategies that explain how or why a particular outcome occurs. Sample outcome measures include accuracy of performance, timeliness of action, number of errors; sample process measures include quality of team communications, accuracy of team Situation Awareness, and adequacy of team leadership. Although successful outcomes are the ultimate goal of team training, the measurement of processes is critical for diagnosing performance problems. Feedback to trainees based

Report Documentation Page

Form Approved
OMB No. 0704-0188

Public reporting burden for the collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to a penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.

1. REPORT DATE **01 JUL 2007**	2. REPORT TYPE **N/A**	3. DATES COVERED **-**
4. TITLE AND SUBTITLE **Team and Collective Performance Measurement**		5a. CONTRACT NUMBER
		5b. GRANT NUMBER
		5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)		5d. PROJECT NUMBER
		5e. TASK NUMBER
		5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **DSTL, Policy and Capability Studies Bedford Technology Park Thurleigh, Bedfordshire UNITED KINGDOM**		8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)		10. SPONSOR/MONITOR'S ACRONYM(S)
		11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release, distribution unlimited

13. SUPPLEMENTARY NOTES
See also ADM002028.

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**UU**	**16**	

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std Z39-18

on outcomes alone may be misleading and detrimental to learning. For example, teams may stumble on the correct decision or course of action despite the use of flawed processes. If feedback is outcome-based, these flawed processes will not be corrected [2].

2.1 Example of Measure of Team Outcomes – UPAS

The US Army Research Institute developed the Unit Performance Assessment System (UPAS) to help eliminate some of the limitations with the feedback capabilities of SIMNET. The UPAS has also been used in a cost-effectiveness evaluation of the Multi-service Distributed Testbed (MTD2) [3].

The system provides students and instructors with timely and useful feedback by performing all statistical analyses in real or near real-time. A UPAS collects and records data packets from SIMNET and translates and organises derived information into a relational database. This information is further manipulated onto map and graphic displays of unit performance that can be used during SIMNET after action reviews. In SIMENT the UPAS collected the following types of Protocol Data Units (PDUs): vehicle appearance, vehicle status, status change, fire, indirect fire and impact (vehicle or ground). The UPAS used five data sources to analyse unit performance in a DIS environment: network data, terrain data, units plans for the operation, radio communication and direct observation of participant behaviour.

2.2 Measures of Team SA and Shared Mental Models

There are two other very important concepts underlying team performance for which measures need to be developed: team Situation Awareness (SA) and shared mental models.

Team SA: SA is important to teams as it allows team members to be attentive to changes in the environment and anticipate the consequences of these variations [4]. A useful definition has been developed for an aviation context: Team SA has been defined as the crew's understanding of flight factors that can have an impact on the mission effectiveness and safety of the crew. Muniz et al. have identified the flight factors and have identified behavioural indicators of low and high team SA [4, 5]. Low team SA includes lack of communication, fixation, deviating from SOPs, violating limitations, using undocumented procedures, etc. Examples of high team SA include confirming information, re-checking of old information, identifying potential problems, noting deviations, having contingency plans, responding quickly to radio messages. Measures are required to evaluate the team SA of participants in a collective training exercise.

Shared Mental Models: For effective team functioning, team members need to be able to predict the needs and information expectations of other team-mates and anticipate actions. This ability is explained by hypothesising that members exercise shared or common knowledge bases, i.e., shared mental models. Shared mental models have been defined as 'Knowledge structures held by members of a team that enable them to form accurate explanations and expectations for the task, and in turn to coordinate their actions and adapt their behaviours to the demands of the task and other team members' [6].

The greater the degree of overlap in team members' models, the greater the likelihood that members will predict, adapt, and co-ordinate with one another successfully. This concept has important implications for scenarios where teams are required to co-ordinate with teams from other services and nations, where the degree of overlap may not be as great as between members from the same units. Measures are required to assess the degree of overlap between participants in a training exercise.

2.3 Example Measure of Team SA – SALIENT

Few methods for measuring team SA exist. This section examines one method known as SALIANT (<u>S</u>ituational <u>A</u>wareness <u>L</u>inked <u>I</u>nstances <u>A</u>dapted to <u>N</u>ovel <u>T</u>asks) which was developed at NAWC [4, 5].

SALIANT is an event-based approach which evaluates teams based on behaviours associated with team SA. It is similar in approach and format to TARGETs; it provides a behavioural checklist and has been found to have high inter-rater reliability. The SALIANT methodology comprises of 5 phases:

Phase 1: Delineation of behaviours theoretically linked to team SA. 21 generic behaviours have been identified from the literature and these have been clustered into 5 categories:

- Demonstrating awareness of surrounding environment;
- Recognising problems;
- Anticipating a need for action;
- Demonstrating knowledge of tasks; and
- Demonstrating awareness of important information.

Phase 2: Development of scenario events to provide opportunities to demonstrate team SA behaviours. These events were based on SME inputs and a team task analysis.

Phase 3: Identification of specific, observable responses. The behavioural indicators were transformed into observable responses based on 5 flight factors identified as crucial for attaining crew situational awareness, i.e., mission objectives, orientation in space, external support equipment status and personal capabilities.

Phase 4: Development of script. To ensure consistency across teams – when events should be introduced, what information to be provided and how to respond to teams.

Phase 5: Development of structured observation form. The form was developed to rate teams on the number of specific observable behaviours exhibited, i.e., coded whether hit or a miss.

2.4 The Role of Mental Models in Team Effectiveness

Although it has been proposed that shared mental models may hold the key for understanding and explaining team performance, there are few methods for investigating shared mental models. Where reports describe the application of certain techniques, the details of how to administer and analyse are sparse.

Based on a survey of existing research in the area of team behaviour and cognition, UK researchers funded by MoD [6] developed a generic theoretical representation of mental model functionality in command planning teams. This representation provided hypotheses for a pilot trial. They also undertook a comprehensive survey of existing data collection and assessment methods. Some of these methods were modified and new ones were developed to capture mental model data and evaluate hypotheses in a pilot study. One of these looked at the representation of mental models in command teams. The representation developed assumes that mental models can be conceived as a network of interrelated models that pass each other results of their processing. These models can be divided in 3 types:

1) Situation assessment;
2) Taskwork including models of task, equipment, time, team, individual, information; and
3) Teamwork including models of enemy plans, situation development, time, movement, combat, enemy capability, own force capability.

Each model represents a view about some aspect of the team's world. Models contain links in to other models. To complicate things further, there are also ***experiential*** and ***dynamic*** forms of mental models.

Experiential models are built up from past experiences and training; dynamic models are formed from an integration of experiential mental models and information derived from current operating environment.

It is assumed that all models must exist somewhere in team, but not all models need to be held by all members of the team. For an experienced team, the SA and teamwork models are likely to be shared to a significant extent, thus fewer requirements for taskwork models to be shared. Not all team members will have the same constructs nor represent them in same form.

To test these hypotheses, researchers developed a pre-exercise interview aimed to capture experiential teamwork mental models. A cluster analysis was conducted on the lists of rated characteristics to produce quadrant graphs for each team. These show where and how team members think similarly or differently on pertinent issues. The graphs include two variables; the level of consensus for a characteristic, i.e., number of individual who think the characteristic is related to effective teamwork; and the level of criticality for a characteristic, i.e., the degree to which they think the characteristic is critical for effective teamwork. Four quadrants were defined:

- High consensus / high criticality: most people believe characteristics critical for effective teamwork.
- High consensus / low criticality: majority consider relate to, but not critical for, effective teamwork.
- Low consensus / high criticality: one or a minority consider very important for teamwork.
- Low consensus / low criticality: one or a minority considers as not very important.

The graphs provided a profile of thinking within teams and highlight the shared perceptions and potential difference between team members.

Post exercise, teamwork analysis methods were designed to supplement the findings concerning behaviours and dynamic model utility observed and as a mechanism of exposing team's shared perceptions if teamwork. Ratings in importance and extent to which team possessing teamwork characteristics were analysed to assess the levels of disparity between team perceptions and the extent to which opinions were shared.

The results showed that teams mentioned very similar characteristics in particular the ones considered to be core to teamwork, e.g., trust and confidence, situation awareness, individual good taskwork knowledge and skills, providing and receiving performance feedback. The researchers found the method quick and simple to use and provided an effective means for analysing a team's perspective in teamwork.

3.0 COMMUNICATION ANALYSIS

Team members cooperating within or between units and teams need to coordinate their actions. This cooperation is mainly mediated by verbal and written communication, and gestures. In the network centric warfare-oriented defence, the need for communication is apparent, as is the need for communication analysis. Team communication factors have proven to be related to team performance [7]. Some areas of interest are:

- Who is communicating with whom?
- What is communicated?
- What is communicated overtly versus implicitly?
- Are the operators explicitly aware of important situation aspects?
- Which media/channels are used?

- Are there problems/errors in the communication?
- Do the problems have serious consequences on the performance?

Communication analysis can be used in addition to task analysis. It can provide information regarding changes in behavior when modifying/upgrading systems. Many of today's VR systems enable logging of both verbal transactions using push-to-talk-buttons, as well as data transfers on, e.g., enemy positions, which can provide a rich data source for post-event analysis. The communication analysis methods often involve using transcriptions of spoken communication for in-depth examination, including analysis of speech frequencies for different categories of communication, problem occurrences, and communication quality ratings. (See also Sections 4 and 5).

3.1 An Example of Communication Analysis – The IOC Model

UK MoD has funded research to develop metrics to quantify the effectiveness of training applicable to all levels and types of Armed Force training. This work has resulted in a novel approach to representing the performance of teams [8], namely the Integration Organisation and Cohesion (IOC) count analysis model. The overall objective of this research is to develop objective metrics and a methodology that will provide the MoD with a quantitative means of representing collective training in high level Operational Analysis (OA), balance of investment and cost-effectiveness models.

A descriptive model has been developed as a framework on which to base the work. This model proposes that the development of collective performance is based on improvements in integration (I), organisation (O) and cohesion (C) across the relevant set of people, i.e., the IOC model. The model is output-based and aims to assess how well a collective is working together and thus can be used to quantify the extent to which collective training has had an impact.

The IOC model breaks down the team's activities into two types: taskwork and teamwork. It is assumed that successful team outcomes rest on both good taskwork (sub-unit, e.g., formation performance) and good teamwork (processes), and that the primary purpose of team training is teaching good teamwork.

The central idea of the model is that there are three patterns of interaction within the teams:

- Actions based on response to orders;
- Actions based on the need to co-ordinate with other entities; and
- Actions based on loyalty to the team.

These can be defined in terms of three constructs:

- **Integration:** the extent to which realignment of goals arises from interventions by the collective leader. Evidence includes orders/commands coming from the leader of the collective, or information flow between the leader and the team.
- **Organisation:** the extent to which the functions of the entities are distributed and aligned to achieve the common goal. Evidence includes lateral communications used to share situational awareness, or make suggestions to each other.
- **Cohesion:** the extent to which realignment of goals arises from the entities themselves. Evidence includes reinforcing/supporting type communications.

The model hypothesises that the definition of the state of the team in terms of Integration, Organisation and Cohesion would provide an indication of how effectively the collective is likely to perform. It is assumed that appropriate scores for these attributes would lead to patterns of behaviour that support the overall goal of the team. Team training then modifies these behaviours in a manner that enhances the likelihood of achieving the team outcome.

In summary, the IOC Count Analysis technique has demonstrated utility for quantifying the value of team training. However, the technique is probably more applicable to teams within the land and naval domains, where the command structure is more hierarchical, and where communication is central to success.

4.0 DISTRIBUTED VR SYSTEMS – EVENT BASED TRAINING

Some tools have been adapted to measure teamwork in a distributed training environment. These tools were developed in the context of an instructional approach known as Event Based Training (EBT) which links learning objectives, exercise events, performance measures and After Action Review (AAR) or debrief.

Basically the EBT approach involves:

- Specification of Training Objectives (TOs): critical tasks, conditions and standards of performance.
- For each TO, the identification of specific learning objectives: these define the specific focus of exercise (we haven’t talked much about learning objectives in past). Learning objectives represent behaviours which have been deficient in the past, are subject to skill decay, or so difficult to perform need frequent practice.
- Identification of “trigger events” for each learning objective- these create opportunity for participants to demonstrate ability to perform tasks associated with learning objectives. They also provide controlled situations in which evaluators can assess performance.
- Development of performance measures used to assess task performance during each event.
- Examination of measurement data and presentation in manner to support feedback.

Dwyer et al. have been involved in the first systematic application of the EBT approach in a distributed training environment [9]. This was used to develop performance measures, namely the TARGET checklist the TOM instrument. These are outlined below, together with a description of how they were used in two case studies.

4.1 The TARGET Checklist

TARGET stands for Targeted Acceptable Responses to Generated Events or Tasks [10]. The method is event-based and involves the identification of events for a training session which serve as triggers for team members to exhibit examples team behaviours.

In addition, for each of these events, acceptable responses (i.e., the TARGETs) are identified in advance of the exercise. Anticipated behaviours are based on training manuals, SOPS, doctrine and SME inputs. Behaviours are then arranged into a checklist in the approximate order they will occur. As the exercise unfolds, observers score each item as acceptable, unacceptable or unobserved. An implicit assumption in the TARGETs methodology is that behaviours are observable and the instructor can determine them as being present, i.e., a “HIT” or absent, i.e., a “MISS”.

Performance can be assessed in number of ways: the proportion of behaviours correctly performed relative to total set of behaviours can be calculated or behaviours can be grouped into functionally related clusters, which can then be examined to see how well team performed in functional areas.

4.2 The Teamwork Observation Measure

Teamwork Observation Measure (TOM) was derived from performance measurement techniques developed under the US Navy’s tactical decision making under stress [5], and aircrew co-ordination training research. The purpose of TOM is to identify performance strengths and weaknesses and to obtain performance ratings on critical dimensions of teamwork.

TOM includes 4 dimensions of teamwork: communication, team co-ordination, situational awareness and team adaptability. Each dimension is then divided into key factors (see Table 7-1).

Table 7-1: TOM Dimensions and Factors

TOM Dimension	**Factors**
Communication	Correct format
	Proper terminology
	Clarity
	Acknowledgements
Team Co-ordination	Synchronisation
	Timely passing of information
	Familiarity with other's jobs
Situational Awareness	Maintenance of big picture
	Identify potential problem areas
	Remain aware of resources available
	Provide information in advance
Team Adaptability	Back-up plans
	Smooth transition to back-up plans
	Quick adjustment to situational changes

Assessors are required to provide specific comments based on observations made to be highlighted as critical points during feedback. Assessors also provide ratings of how well participants interacted with each other on each of the four teamwork dimensions.

5.0 COLLECTIVE PERFORMANCE ASSESSMENT

The need to assess and measure performance at a collective[1] level presents researchers with a number of challenges. A collective operates at a higher level than a team and involves different roles co-ordinating their activities, without necessarily being co-located and without necessarily having a single recognised leader or identical goals. Certain skills that are important for teams, e.g., communication, co-ordination and information sharing are also key to collective success. However, in a collective there is less likelihood of shared expectations derived from previous experience and reduced area of overlap in shared mental models compared to an established team [11].

To use an example from the air domain, Collective air mission training may involve many aircraft, fulfilling different roles, some directly involved in a mission and some providing support. For example, a 4-ship in an Air to Ground role needing to co-ordinate with Air-to-Air assets, Suppression of Enemy Air Defence (SEAD) assets and Airborne Warning and Control System (AWACS) aircraft. It is the inter-team

[1] 'Collective mission training' is defined as two or more teams training to interoperate in an environment defined by a common set of collective mission training objectives, where each team fulfils a different military role. NATO SAS-013 Study.

rather than intra-team interactions and co-ordination that are important. High level cognitive skills, such as the ability to build and maintain situation awareness or to make tactical decisions in a complex and highly dynamic environment are crucial.

5.1 Implications for Collective Training Assessment Techniques

An understanding of the benefits gained from current collective air training gives an insight into what needs to be captured by training assessment and performance measurement techniques. There is a need for techniques that do not simply capture mission outcomes, but more importantly the underlying cognitive processes and strategies. To truly quantify the training value of collective air training, there is a need to capture some of the less tangible benefits for example positive changes in aircrew's understanding, situational awareness, flexibility and confidence.

Any techniques identified should ideally be of utility in the live environment. For example, whilst observers are able to make valid, albeit subjective judgements of performance and use these to give feedback and guidance to participants, live collective exercises could benefit from a more formal approach. In addition, if techniques could be applied to both live and VR exercises this would enable comparisons of the relative value of both training environments to be made.

5.2 Collective Performance Assessment and Mission Phases

Within the UK, under the sponsorship of the MoD, a programme of applied research has been undertaken to explore the benefits to be gained from using networks of simulators within a VR environment for aircrew collective mission training. Use of networked simulation in this context (in the UK) has become known as UK Mission Training through Distributed Simulation (MTDS) [12]. The approach adopted by UK MTDS researchers advocates a subjective assessment of performance during all phases of the training event [13]. Typically these phases are plan, brief, mission execution and After Action Review (AAR) or debrief. This work has led to the devolvement of tool designed specifically to assess collective performance during all mission phases; the Collective Assessment performance Tool (C-PAT) [15].

C-PAT is being developed by the Defence Science and Technology Laboratory (Dstl), part of the UK MoD. It forms part of an evolving concept of analysis for the UK Mission Training through Distributed Situation (MTDS) initiative and has already demonstrated great utility in providing measures of effectiveness for synthetic collective training. Essentially C-PAT is a 'family' of surveys (listed below in Table 7-2.) designed specifically to facilitate Subject Matter Expert (SME) assessment of collective performance of aircrew throughout all mission phases. Typically these SMEs also undertake the White Force role during both live and virtual collective training events.

Table 7-2: The C-PAT Familiy of Surveys

C-PAT Survey Element	Description
Planning Phase Assessment	WF evaluation of the 'quality' of co-ordination during the planning process on each mission day is an important component of this assessment. This is something that the WF are well used to judging during live collective training exercises. At the end of the planning phase of each mission the WF team were asked to complete a planning assessment questionnaire, giving their expert judgement in areas such as leadership, use of information, time management, thinking about the 'Big Picture', decision making.
Mass Brief Assessment	WF will evaluate the 'quality' of the briefs in terms of clarity, accuracy, big picture information, etc. This survey is still under development.
Assessment Criteria	At the end of each mission the WF are asked to complete an "Assessment Criteria" questionnaire, which asked for assessments on 31 criteria to form a picture of how well a collective exercise is proceeding. Typical criteria are: How effective were the tactics employed during the mission? How appropriate was any review of tactics made as a result of lessons learned? To what extent were the overall objectives of the mission achieved? Were relevant lessons learned and actions thoroughly debriefed?
Mass Debrief Assessment	WF will evaluate the 'quality' of the debrief in terms of clarity, accuracy, and lessons identified. This survey is still under development.
Training Objectives	Participants will be asked to rate to what level the training objectives were supported during the training event. These comprise a number of sub-elements, all of which are given a rating. Scores will then be consolidated to give an overall rating for each of the TOs.
Interoperability	Trust is a vital component of interoperability. One of the benefits of collocation is that it appears to help engender trust in away that may not be possible with distributed players. This survey is still under development.

The C-PAT has been developed on the premise that effective collective processes can really only be assessed by an SME with the appropriate level of domain specific knowledge. The thought processes used in making these judgements are often difficult to articulate and considerable effort has been expended in trying to elicit these from tactical/training experts from the Air Warfare Centre (AWC). The tools are continually being refined with inputs from the AWC, and it is hoped that their involvement in the design of C-PAT, will ensure that these are formulated and worded in a manner that will be understood by end users. The ultimate aim is to develop robust metrics that can be utilised to measure the effectiveness of both live and synthetic collective air training exercises, thus enabling the value of UK MTDS training exercises to be quantified.

At the centre of the C-PAC, are collective performance indicators; these have been derived from benefits identified by participants in live collective exercises. Typical collective performance indicators that have been used to assess the utility of a virtual environment to support mission training are presented in Table 7-3.

Table 7-3: List of Typical Collective Performance Indicators

No.	Collective Performance Competency/Indicator
1	Understanding of own team's role and capabilities
2	Understanding of other team's role and capabilities
3	Understanding of where own team fits into the 'bigger picture'
4	Ability to balance risks – exploring the 'what ifs' of the training scenarios
5	Ability to cope with the 'fog of war'
6	Awareness of the tactical situation (multi-level SA)
7	Within role communication and co-ordination skills
8	Between role communication and co-ordination skills
9	Tactical skills
10	Tactics development
11	Utilisation of role specific skills within the collective environment
12	Ability to understand and implement briefed operational procedures
13	Effectiveness in Commander role
14	Decision making
15	Fluidity in a variety of dynamic situations
16	Confidence in own capabilities
17	Confidence in own team's capabilities
18	Confidence in other teams' capabilities

The C-PAT is still evolving. One area requiring further investigation is measurement of aircrew Situation Awareness particularly their awareness of other team member's roles and intentions. Good Situational Awareness is integral to an effective mission execution phase, but it is difficult to quantify. With regard to the surveys themselves, feedback indicates that aircrew may find it difficult to equate their established rating scales with the required percentage responses. The use of anchored rating scales is to be investigated. However, this is not necessarily a simple solution, as ease of use does not necessarily equate to more meaningful data. Recently a mapping exercise was undertaken between assessment criteria and collective training competencies. Understanding these relationships will further help with quantifying training effectiveness.

Data collection and analysis can be time-consuming when carried out manually. One of the future aspirations for the technique is to provide a rapid and reliable measure of effectiveness of UK MTDS training events. To this end, there are plans to administer surveys in an electronic format. This should also permit the automatic data collection of responses in quantifiable terms.

6.0 OBJECTIVE MEASURES

Objective measures are less debatable than subjective measures, but can lack in contextual value. Objective measures can serve as a basis for comparison with subjective measures to reflect whether attitudes reflect what actually happened during the mission. It is important to develop a robust set of objective measures for a more rigorous assessment to performance and to maximise the benefits of the AAR/debrief session. Some form of data logger is thus an important component of overall the VR system.

The logger should log all data that is generated within an exercise or event. For example within a networked VR training event, data is typically output onto the network in the form of DIS Protocol Data Units – (PDUs) that are generated. All PDUs are time stamped with their time of reception at the logger. The logs provided by the logger can then be replayed during debrief at normal speed, slower then normal speed or faster than normal speed to enable the instructor, exercise director or trainee to fast-forward and pause at a critical mission incident and engage the training audience in further discussion and capture lessons learnt.

Information captured on the data logger will also provide valuable insight as to the health of the system and the integrity of the technical and tactical networks. More importantly it will provide measures of individual, team and ultimately collective performance which can be used to aid debriefing and performance assessment on a number of different axes.

In order to be able to make such assessments it may be necessary to have a baseline against which actual performance during the training event could be measured. For example, the air defenders performance in Weapons Engagement Zone management and control and how they 'pushed the bubble' could be assessed by comparing it with the baseline parameters; speed, height, sensor information, tactical manoeuvres, etc. Objective assessment is a key to a successful AAR and debrief. Significant progress has been made in this area in recent years and a number of bespoke solutions developed to capture objective data necessary to support a more robust evaluation of performance in a virtual training environment. An example is the work undertaken by the Air Force Research Laboratory in Mesa, US as part of their research into Distributed Mission Training (DMT). AFRL has developed a software tool known as PETS (Performance Evaluation Tracking System). [15]. PETS is capable of capturing the objective data necessary to support a robust and real-time evaluation of performance in a DMT training event. Data is organised at several levels to aid assessment. They include RT graphical displays, performance effectiveness learning curves, and statistical analysis at scenario or shot level.

7.0 INDIVIDUAL PERFORMANCE

Whist the chapter has focused on team and collective performance measures, for completeness some examples of individual performance measures are also included. The measures discussed in this section have been developed by the Swedish Defense Research Agency (FOI) and focus primarily on pilot performance and include both subjective and objective assessment techniques.

7.1 FOI Approach to Performance Measurement

FOI has a long tradition of measuring operative performance. Though varying regarding the specific measures, the general approach has always been the combination of subjective measures (e.g., questionnaires, rating scales), objective measures (e.g., data logging), and psycho-physiological measures (e.g., HRV, EPOG). Since the ambition is to use measures that reflect the dynamics of the situation, attempts to reduce the wide range of variables are necessary. The tradition is to use factor analysis for identification of significant compounded indicators. Linear causal model analyses are then performed by means of structural equation modelling (SEM), for example LISREL [16], to test the validity of different causal flow models possible.

The method of assessing performance that is most commonly used by FOI is a modified version of the Bedford Rating Scale [17]. The pilots answer questions using a 10-point scale. The modified scale can be formulated in either first person or third person. It can also be used pseudo-dynamically, that is, that the scale is being used repeatedly, after important aspects, throughout a mission. The measure has been used in several studies [18, 19].

There are sometimes a difference between pilot ratings and instructor ratings. These differences can be explained by different understanding of what constitutes performance. The ratings has shown correlations with Mental Workload (r = -0.55), Situational Awareness(r = 0.52), and Heart Rate (r = – 0.59) [20].

7.2 The FOI Pilot Performance Scale

The FOI Pilot Performance Scale (FOI PPS) is useful for addressing aspects of difficulty, performance, mental capacity, mental effort, information load, situational awareness, and mental workload. The six dimensions are extracted by means of factor analysis and the number of markers range from 3 to 7. The reliability of the dimensions or indices has been tested by means of Cronbach's alpha and they have been cross-validated. The Swedish questionnaire has not been validated in English. The questions are developed to fit in military fixed wing scenarios and relate to flown missions with specific as well as general questions. Relationships between the indices have been analyzed by means of structural equation modeling [21, 22]. Subjects answer by scoring on a 7-point bipolar scale. This measure has been used in several studies and the reliability ranges from 0.73 to 0.90. Indices change significantly as a function of mission complexity.

The FOI PPS significantly relates to psycho-physiological indices such as heart rate and eye point of gaze changes and it correlates 0.79 with mission/task difficulty level, 0.84 with the NASA-TLX and 0.69 with the Bedford Rating Scale. FOI PPS has mainly been used in training simulators and after missions in real aircraft. FOI PPS is not available in English. Examples of (translated) questions are:

- How complex did you find the mission?
- Did you feel forced to disregard or cancel some of your tasks in order to perform optimally on critical tasks?
- To what extent did you feel disturbed by non-critical information?
- Did you have problems monitoring the information on the Tactical Situation Display (TSD)?

The instrument has 6 dimensions: Operative Performance (r =0.74), Situational Awareness (r =0.80), Pilot Mental Workload (r =0.87), Mental Capacity (r =0.77), Information Handling Tactical Situation Display (r =0.92), and Information Handling Tactical Information Display (TI) (r =0.93).

It takes about 5 minutes to answer the questionnaire. Some subjects find the questionnaire too long and time-consuming. The indices are suitable to use in causal analyses [16].

7.3 Objective Measures of Individual Performance

Task performance measures vary between research groups and research areas. Speed and accuracy are used by most research teams in one way or another. The choice of measure is of course dependent on research area (e.g., visual and audio perception), but also on possibilities in the actual situation. Both controlled laboratory experiments and field studies are of interest, and often complement each other in seeking for new solutions. Below follows some examples of dependent measures of performance commonly used.

Angle Estimation (visual perception): Comparisons between targets with or without monocular depth cues (drop-lines) can be used for evaluating different display settings. Subjects perform angle estimations in a 3D virtual environment where the task is to detect a threat and to estimate the angle of a prioritized target in 3D space [23]. Answer is given by pointing a virtual arrow in the estimated direction, from ownship in direction to target. Both azimuth and elevation are measured and analyzed separately, but they can also be analyzed together. Comparisons between angles are also possible, even though the main interest have been to compare with and without additional monocular depth cues.

- Angle Estimation (audio): Subjects perform angle estimations from an audio signal (speech and noise) with the intent of comparing two 3D-audio systems spatial resolution. One system is an expensive professional hardware solution and the other an inexpensive software application, further developed at FOI.
- Relative Height Estimation: Relative height estimation can be used in a flight situation when evaluating the effect of using monocular depth cues (drop-lines and cone attached to the ground or to a fixed plane). The subject's task is to estimate which target symbol is closest or most distant compared to own ship [24]. One important point using relative estimation is that the measure is non-metric (compared to angle estimation), which can be of importance when using dependent measures in a three-dimensional virtual setting. According to some researchers [25], a 3D virtual environment will create different errors in x, y and z-axis. This kind of problem speaks for the use of non-metric measures in a 3D setting.
- Future Collision Point: In the flight domain future collision points or risk of collisions [26] are of great importance. The subject's task is to select which of a number of the targets has a collision course with the own aircraft. Both speed (Response Time) and accuracy can be measured.
- Deviation from Flight Path: Can be used as a performance measure when flying, e.g., 'tunnel in the sky' or as a secondary dependent measure when performing another task.
- Relative Size Estimation: Can be used when comparing settings were monocular and stereoscopic vision is in focus, including different techniques for stereo presentation and other VE techniques like tactile displays.
- Color Discrimination – Staircase Method to find Just Noticeable Differences (JND): Color perception can be affected during different g-loads during high performance flight. One method is a staircase method with different colors. The baseline for JND at some well known colors is known [27] and can be compared with the JND values acquired in a centrifuge setting.
- Color Identification: Pilots performed identification of well known colors during g-load.
- Symbol Identification: Aircraft vibrates at different amplitudes and frequencies that might cause problems reading text or understanding symbols. To understand the effect of frequencies and amplitudes, experiments were conducted with vibrating symbols at different frequencies [28]. Symbol identification can also be used to evaluate if g-load affects identification during modest or high g-load.
- Balance Measures: Investigation of visual flow effectiveness includes studies of display effects of visual vertical variation on observer balance. Thereby the impact on perceived spatial orientation was studied, with greater postural sway linked to increased proprioceptive and vestibular suppression. Thus, greater postural sway reflects increased display effectiveness [29].
- Reaction Time: Often used as a performance measure in combination with measures of accuracy.
- Alarm Sound Categorization: Subjects performed categorizations of different alarm sounds together with estimated vigilance, duration, and audibility.

8.0 SUMMARY

This chapter has focused primarily on team and collective performance measures. Whilst it has provided some examples of team and collective measures, it should be noted there exists a rich source of information in literature concerned with performance measurement and human factors. Two recent publications which are recommend to readers interested in this topic are; Performance Measurement – Current Perspectives and Future Challenges', edited by Winston Bennett, Charles Lance and David Woehr, published by LEA, 2006 and Human Factors Methods – A Practical Guide for Engineering and Design, by the Human Factors Integration Defence Technology Centre, published by Ashgate, 2005.

9.0 REFERENCES

[1] Naval Air Warfare Center, 'Performance measurement in teams', Training Series No. 2.

[2] Smith-Jentsch, K.A., Johnston, J.H. and Payne, S.C. 'Measuring team-related expertise in complex environments'. In J.A. Cannon-Bowers and E. Salas (Eds.). 'Making decisions under stress: implications for individual and team training'. 1998. APA Press.

[3] Taylor, H.L., Orlansky, J., Levine, D.B. and Honig, J.G. 'Evaluation of the performance and cost-effectiveness of the Multi-Service Distributed Training Testbed (MDT2)'. Proceedings of Royal Aeronautical Society Conference, November 96.

[4] Muniz, E., Bowers, C.A., Stout, R.J. and Salas, E. 'The validation of a team situational measure'. Proceedings of the 3rd Annual Symposium and Exhibition on Situational Awareness in the Tactical Environment, Patuxent River, MD, 62, 1998.

[5] Muniz, E.J., Stout, R.J., Bowers, C.A. and Salas, E. 'A methodology for measuring team Situational Awareness: Situational Awareness Linked Indicators Adapted to Novel Tasks (SALIANT)'. Proceedings of Symposium on "Collaborative Crew Performance in Complex Operational Systems", UK, 20-22 April 1998.

[6] Pascual, R.G., Henderson, S.M. and Mills, M.C. 'Understanding and supporting team cognition-Year 1 report and recommendations', DERA/CHS/MID/CR980122/1.0, June 1998.

[7] Sexton, J.B. and Helmreich, R.L. 'Analyzing cockpit communication: The links between language, performance, error and workload'. Paper presented at the Proceedings of the 10th International Symposium on Aviation Psychology, Columbus, OH. 1999.

[8] Belyavin, A. 'Quantify the effectiveness of collective training: Final Technical Summary Phase I'. QINETIQ/KI/CHS/CR033077. December 2003.

[9] Dwyer, D.J., Randall, L.O., Salas, E. and Fowlkes, J.E. 'Performance measurement in distributed environments: initial results and implications for training'.

[10] Dwyer, D.J., Oser, R.L., Fowlkes, J.E. and Lane, N.E. 'Team performance measurement in distributed environments: the TARGETs methodology' In M.T. Brannick, E. Salas and C. Prince (Eds.). Team performance assessment and measurement: theory, methods and applications. 1997.

[11] McIntyre, H.M. and Smith, E. 'Collective operational aircrew training research: Progress report FY99 and FY00'. QinetiQ/FST/CSS/TR011124/1.0, November 2001.

[12] Smith, E. 'Final Report on UK MTDS Study Part 2 – Detailed Full Report'. Dstl/CR07001V1. September 2003.

[13] Smith E. and McIntyre, H. 'Simulation and collective training – creating a total experience'. The Royal Aeronautical Society, Flight Simulation DERA/AS/FMC/CP010229. June 2000.

[14] Smith, E. 'Quantitative Techniques to Assess the Value of Training Output'. Dstl/CR17059 v1.0. September 2005.

[15] Gehr, S.E., Schreiber, B. and Bennett, W. 'Within-Simulator Training Effectiveness Evaluation'. Proceedings of Interservice/Industry Training, Simulation and Education Conference (I/ITSEC) 2004.

[16] Jöreskog, K.G. and Sörbom, D. 'LISREL8: Structural equation modeling with the SIMPLIS command language'. Hillsdale: Lawrence Erlbaum Associates. 1993.

[17] Berggren, P. (2000). 'Situational awareness, mental workload, and pilot performance – relationships and conceptual aspects'. Linköping: Human Sciences Division. FOA-R-00-01438-706-SE.

[18] Svensson, E., Angelborg-Thanderz, M., Sjöberg, L. and Olsson, S. (1997). 'Information complexity-mental workload and performance in combat aircraft'. Ergonomics, 40, No. 3, pp. 362-380.

[19] Borgvall, J., Nählinder, S. and Andersson, J. (2002). WVR-Illustrator Evaluation: 'Using Pilot Expertise for Future Development' (Scientific No. FOI-R--0710--SE). Linköping, SE: The Department for Man-System Interaction at the Swedish Defense Research Agency.

[20] Magnusson, S. and Berggren, P. (2001). 'Measuring pilot mental status', NAMA conference, Stockholm.

[21] Svensson, E., Angelborg-Thanderz, M. and Wilson, G.F. (1999). 'Models of pilot performance for systems and mission evaluation – psychological and psycho-physiological aspects'. AFRL-HE-WP-TR-1999-0215.

[22] Svensson, E. and Wilson, G.F. (2002) 'Psychological and psycho-physiological models of pilot performance for systems development and mission evaluation'. International Journal of Aviation Psychology, Vol. 12 (1).

[23] Andersson, P. and Alm, T. (2003). 'Perception Aspects on Perspective Aircraft Displays'. Displays, 24, 1-13.

[24] Lif, P. and Alm, T. (2004). 'Relative Height in 3D Aircraft Displays'. Human Performance, Situation Awareness and Automation Technology Conference.

[25] Smallman, H.S., John, M.S. and Cowen, M.B. (2002). 'Use and misuse of linear perspective in the perceptual reconstruction of 3-D perspective view displays'. Paper presented at the Human Factors and Ergonomics 46th Annual Meeting, Baltimore, Maryland, US.

[26] Endsley, M.R. (1995). 'Toward a Theory of Situation Awareness in Dynamic Systems'. Human Factors, 37(1), 32-64.

[27] MacAdam, D.L. (1942). 'Visual sensitivities to color differences in daylight'. Journal of the Optical Society of America, 32(5), 247-274.

[28] Andersson, P. and Hofsten, C.V. (1999). 'Readability of vertically vibrating aircraft Displays'. Displays, 20, 23-30.

[29] Eriksson, L. and von Hofsten, C. (2004). 'Effects of visual flow display of flight maneuvers on perceived spatial orientation'. Human Factors. (Accepted manuscript.)